BÖLÜM 7

MODEL GELİŞTİRME: DEĞİŞKEN SEÇİMİ

GİRİŞ

- Önceki üniteler, en küçük kareler regresyonunun hesaplaması ve yorumlanması ile ilgiliydi.

- Bölüm 5'teki vaka çalışması dışında, modelde bağımsız değişkenlerin kullanılması gerektiği ve hangi gösterimde belirtildiğinin bilindiği varsayılmıştır.

- En küçük kareler tahmincilerinin özellikleri modelin doğru olduğu varsayımına dayanmaktadır.

REGRESYON EŞİTLİĞİ KULLANIM ALANLARI

En küçük kareler analizinin amacı – regresyon eşitliğinin nasıl kullanılacağı - modelin nasıl oluşturulacağı biçimini etkileyecektir. Hocking (1976), Mallows (1973b)'da yer alan regresyon eşitliklerinin potansiyel kullanımını şu şekilde ilişkilendirmektedir:

1. Tepki değişkeninin davranışının iyi bir açıklamasını sunmak;
2. Gelecekteki tepkileri tahmin etmek ve ortalama tepkileri tahmin etmek;
3. Extrapolasyon ya da veri seti aralığı dışındaki tepkileri tahmin etmek;
4. Parametre tahminleri;
5. Değişen seviyede girdilerle sürecin kontrolü ve
6. Sürecin gerçekçi bir modelini geliştirmek.

EN KÜÇÜK KARELER ÜZERİNE DEĞİŞKEN SEÇİMİNİN ETKİLERİ

- Değişken seçiminin en küçük kareler sonuçları üzerindeki etkileri örtük olarak seçimin cari veri setindeki bilgiye dayanmadığı durum için geliştirilmiştir.

- Bu genellikle değişken seçimi tekniklerinin tartışıldığı önceki bölümde olduğu gibi pek gözlenmez.

- Fakat bu durum için teorik sonuçlar değişken seçimi ile ilgili motivasyon sunmaktadır.

TÜM OLASI REGRESYONLAR

- Veri setindeki bağımsız değişkenler dik ise, tasarımlı deneyde olduğu gibi, her bir değişken için en küçük kareler sonuçları hangi diğer değişkenlerin modelde olduğundan bağımsız olarak aynı kalmaktadır.

- Bu durumlarda tek en küçük kareler analizi sonuçları bağımsız değişkenleri modelde tutmak için seçimde kullanılabilir.

- Genelde, halbuki, bağımsız değişkenler dik olmayacaktır.

- Dik olmama gözlemsel veri seti ile beklenmelidir ve farkedilemeyen talihsizlikler nedeniyle tasarımlı deneylerde sıklıkla ortaya çıkacaktır.

TABLO 7.1. *Özet istatistik R^2, MS(Res), R^2_{adj} ve C_p, Linthurst veri seti için olası tüm regresyonlarda, SALINITY, pH, K, Na, Zn bağımsız değişkenleri kullanılır. Tüm modeller sabit terim içermektedir. (Veri seti izin ile elde edilmiştir.)*

p'	*Değişkenler*	R^2	*MS(Res)*	R^2_{adj}	C_p	*AIC*	*SBC*
2	*pH*	.599	178618	.590	7.4	546.1	549.8
	Zn	.390	272011	.376	32.7	565.1	568.7
	Na	.074	412835	.053	70.9	583.8	587.5
	K	.042	427165	.020	74.8	585.4	589.0
	SAL	.011	441091	−.012	78.6	586.8	590.4
3	*pH, Na*	.658	155909	.642	2.3	541.0	546.4
	pH, K	.648	160865	.631	3.6	542.2	547.8
	pH, Zn	.608	178801	.590	8.3	547.1	552.5
	SAL, pH	.603	181030	.585	8.9	547.7	553.1
	SAL, Zn	.553	204209	.531	15.1	553.1	558.5
	Na, Zn	.430	260164	.403	29.9	564.0	569.4
	K, Zn	.415	266932	.387	31.7	565.2	570.6
	SAL, Na	.078	421031	.034	72.5	585.7	591.1
	K, Na	.074	422520	.030	72.9	585.8	591.2
	SAL, K	.053	432069	.008	75.4	586.8	592.3
4	*pH, Na, Zn*	.663	157833	.638	3.8	542.4	549.7
	pH, K, Na	.660	158811	.636	4.1	542.7	549.9
	SAL, pH, Na	.659	159424	.634	4.2	542.9	550.1
	SAL, pH, K	.652	162636	.627	5.0	543.8	551.0
	pH, K, Zn	.652	162677	.627	5.1	543.8	551.0
	SAL, pH, Zn	.637	169900	.610	6.9	545.7	553.0
	SAL, K, Zn	.577	198026	.546	14.2	552.6	559.9
	SAL, Na, Zn	.564	203666	.533	15.6	553.9	561.1
	K, Na, Zn	.430	266509	.388	31.9	566.0	573.2
	SAL, K, Na	.078	431296	.010	74.5	587.7	594.9
5	*SAL, pH, K, Zn*	.675	155832	.642	4.3	542.7	551.8
	SAL, pH, Na, Zn	.672	157312	.639	4.7	543.2	552.2
	pH, K, Na, Zn	.664	160955	.631	5.6	544.2	553.2
	SAL, pH, K, Na	.662	162137	.628	5.9	544.5	553.6
	SAL, K, Na, Zn	.577	202589	.535	16.1	554.6	563.6
6	*SAL, pH, K, Na, Zn*	.677	158622	.636	6	544.4	555.2

ADIMSAL REGRESYON YÖNTEMLERİ

- Olası tüm regresyonlar için gerekli hesaplamalardan daha az bir hesaplama gerektiren alternatif değişken seçimi yöntemleri geliştirilmiştir.
- Bu yöntemler iyi alt küme modellerini belirlemektedir ancak en iyisi değildir.
- Bu yöntemler adımsal regresyon yöntemleri olarak adlandırılır.
- Alt küme modelleri artık kareler toplamı üzerine en büyük etkiyi yapacak bir değişkeni sırasıyla ekleme ya da eleme belirlenmektedir.
- Bu adımsal yöntemler her bir alt küme büyüklüğü için en iyi alt kümeyi bulmayı garanti etmemektedir.
- Ayrıca farklı yöntemler kullanılarak elde edilen sonuçlar birbiriyle örtüşmeyebilir.

TABLO 7.2. *Linthurst veri seti kullanılarak yapılan ileriye doğru değişk seçimi için özet istatistikler, SLE = .50.*

Adım	*Değişken*	*Kısmi SS*	*MS(Res)*	R^2	F^a	*Prob* > F^b
1. En iyi tek değişkeni belirle ve giriş için test et:						
	Sal	204048	441091	.0106	.46	.5001
	pH	11490388	178618	.5994	64.33	.0001
	K	802872	427165	.0419	1.88	.1775
	Na	1419069	412834	.0740	3.44	.0706
	Zn	7474474	272011	.3899	27.48	.0001
	En iyi 1-değişkenli model: *pH*				$C_p = 7.42$	
2. En iyi ikinci değişkeni belirle ve giriş için test et:						
	Sal	77327	181030	.6034	.43	.5170
	K	924266	160865	.6476	5.75	.0211
	Na	1132401	155909	.6584	7.26	.0101
	Zn	170933	178801	.6083	.96	.3338
	En iyi 2-değişkenli model: *pH Na*				$C_p = 2.28$	
3. En iyi üçüncü değişkeni belirle ve giriş için test et:						
	Sal	11778	159424	.6590	.07	.7871
	K	36938	158804	.6604	.23	.6322
	Zn	77026	157833	.6625	.49	.4888
	En iyi 3-değişkenli model: *pH Na Zn*				$C_p = 3.80$	
4. En iyi dördüncü değişkeni belirle ve giriş için test et:						
	SAL	178674	157312	.6718	1.136	.2929
	K	32964	160955	.6642	.205	.6533
	En iyi 4-değişkenli model: *pH Na Zn SAL*				$Cp = 4.67$	
5. Giriş için sonuncu değişkeni test et:						
	K	106211	158622	.6773	.670	.4182
	SLE = .50 sonuncu değişken eklenmiştir				$C_p = 6.00$	

TABLO 7.3. *Linthurst veri seti kullanılarak geçmişe dönük eleme için özet istatistikler, SLS = .10. Tüm modeller sabit içerir.*

Adım	*Değişken*	*Kısmi SS*	R^{2a}	F^b	$Prob > F^c$
0	*Model* : tüm değişkenler $R^2 = .6773$, $C_p = 6$, $s^2 = 158,616$ with 39 d.f.				
	SAL	251,921	.6642	1.59	.2151
	pH	1,917,306	.5773	12.09	.0013
	K	106,211	.6718	.67	.4182
	Na	46,011	.6749	.30	.5893
	Zn	299,209	.6617	1.89	.1775
1	*Model* : Na çıkarıldı; $R^2 = .6749$, $C_p = 4.30$, $s^2 = 155,824$ with 40 d.f.				
	Sal	436,496	.6521	2.80	.1020
	pH	1,885,805	.5765	12.10	.0012
	K	732,606	.6366	4.70	.0361
	Zn	434,796	.6522	2.79	.1027
2	*Model* : Zn çıkarıldı; $R^2 = .6522$, $C_p = 5.04$, $s^2 = 162,636$ with 41 d.f.				
	Sal	88,239	.6476	.54	.4656
	pH	11,478,835	.0534	70.58	.0001
	K	935,178	.6034	5.75	.0211
3	*Model* : Sal çıkarıldı; $R^2 = .6476$, $C_p = 3.59$, $s^2 = 160,865$ with 42 d.f.				
	pH	11,611,782	.0419	72.18	.0001
	K	924,266	.5994	5.75	.0211

DUR. Eğer kalan herbir değişken için Prob> F ,SLS = .10 u aşarsa DUR. Nihai model pH, K ve sabit terimi içerir.

ALT KÜME BÜYÜKLÜĞÜ SEÇİMİ İÇİN KRİTER

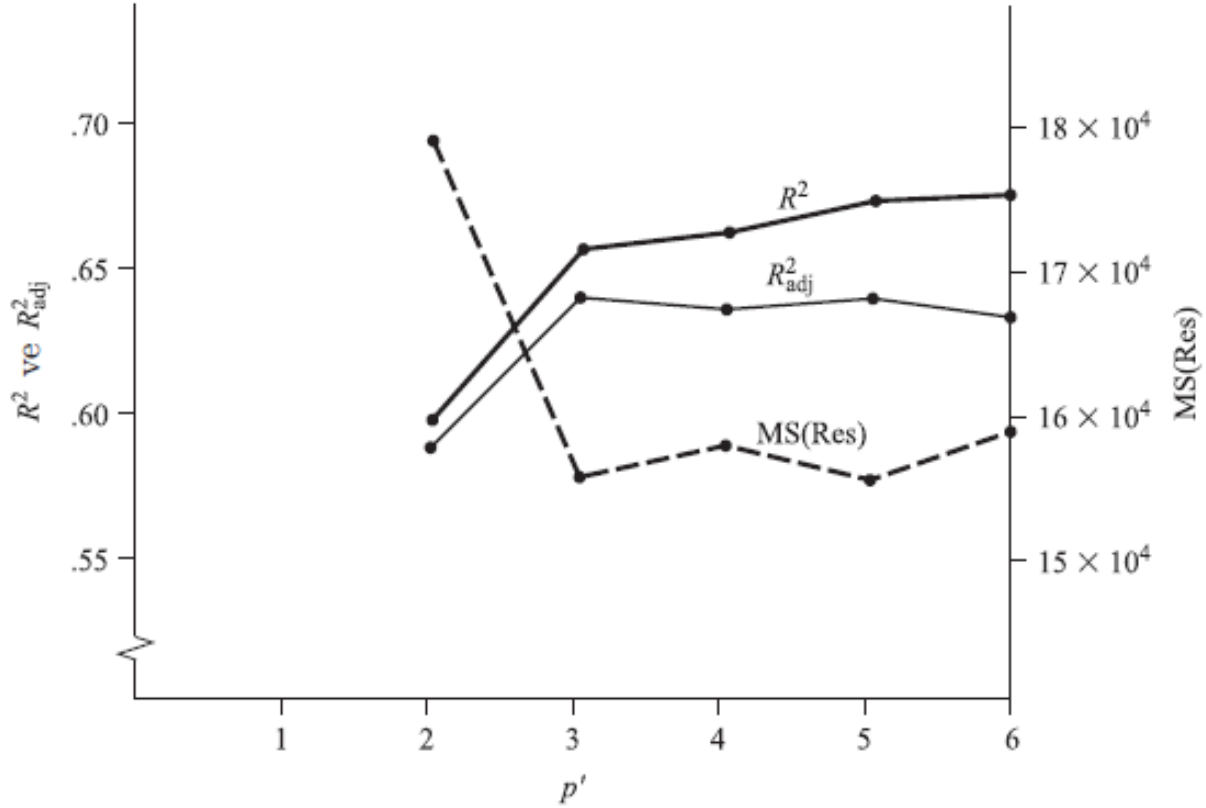
- Alt küme büyüklüğü ile ilgili bir çok seçim kriteri önerilmiştir.
- Bu kriterler küçük artık kareler toplamına dayanan, gerektiğince az parametre içeren tutumluluk ilkesine dayanmaktadır.
- Hocking (1978) 8 durdurma kuralını incelemektedir, Bendel ve Afifi (1977) 8 tane (tamamı Hocking'ine ile aynı değildir) geleceğe dönük kriteri karşılaştırmakta ve PROC REG'deki RSQUARE yöntemi 12 adet seçme opsiyonu sunmaktadır.

KATSAYI BELİRLEME

Belirleme katsayısı R^2 modelde bağımsız değişkenler tarafından açıklanan bağımlı değişkenin toplam (düzeltilmiş) kareler toplamını göstermektedir:

$$R^2 = \frac{\text{SS(Regr)}}{\text{SS(Toplam)}}. \qquad (7.1)$$

R^2 nin davranışı



ŞEKİL 7.1. *R^2, R^2_{adj} ve MS(Res) Linthurst veriseti için her bir alt küme büyüklüğü ile en iyi modelde ele edilen p'ye karşı çizilmiştir.*

ARTIK KARELER ORTALAMASI

Artık kareler ortalaması MS(Res) eğer model tüm ilgili bağımsız değişkenleri içeriyorsa σ^2 nin tahminidir. Eğer ilgili bağımsız değişkenler ihmal edilmişse, kalıntı ortalama kare yukarı yönlü yanlı olacaktır. Önemsiz bağımsız değişken ilave etmek artık kareler ortalaması üzerinde az bir etkiye sahip olacaktır. Böylece artık kareler ortalamasının beklenen davranışı, değişkenler modele ilave edildikçe, σ^2 ye doğru azalacaktır. Ayrıca tüm ilgili değişkenler modele ilave edildiğinde σ^2 etrafında dalgalanacaktır.

UYARLANMIŞ BELİRLEME KATSAYISI

Uyarlanmış R^2, R^2_{adj} ile gösterilmekte ve R^2 nin bağımsızlık derecesi ile yeniden ölçeklendirilmesidir. Böylece kalıntı kareler toplamı yerine ortalama kareler oranını içerir.

$$\begin{aligned} R^2_{adj} &= 1 - \frac{\text{MS(Res)}}{\text{MS(Toplam)}} \\ &= 1 - \frac{(1-R^2)(n-1)}{(n-p')}. \end{aligned} \tag{7.2}$$

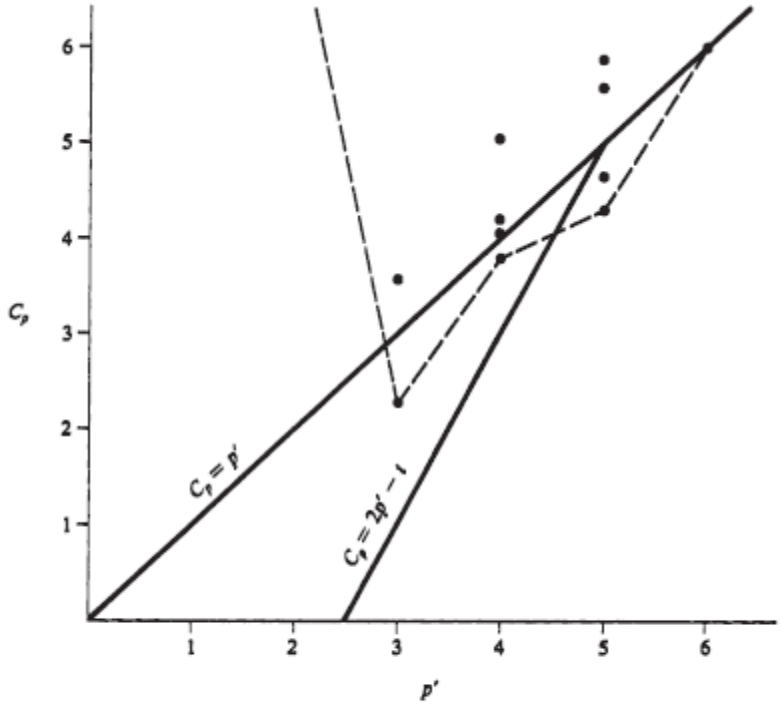
Bu gösterim bağımsızlık derecesinin etkisini ortadan kaldırır ve R^2den daha karşılaştırılabilir bir miktar verir. R^2 den farklı olarak, R^2_{adj} nin değişkenler modele eklendikçe artması gerekmemektedir. R^2_{adj} değeri değişkenler eklendikçe üst limit etrafında istikrara kavuşacaktır. Bu üst limite yakın en basit R^2_{adj} en iyi model olarak seçilmektedir. R^2_{adj} MS(Res) e yakından ilişkilidir (bk. Eşitlik 7.2) ve aynı sonuçlara yol açacaktır.

MALLOW'UN Cp İSTATİSTİĞİ

C_p istatistiği cari veri seti tahmini için stadart hale getirilmiş toplam ortalama karelerin tahminidir (Hocking, 1976). C_p istatistiği ve C_p grafiği başlangıçta Mallows tarafından tanımlanmıştır [daha önceki referanslar için bk. Mallows (1976a)]. C_p istatistiği şu şekilde hesaplanmaktadır:

$$C_p = \frac{\text{SS(Res)}_p}{s^2} + 2p' - n, \tag{7.3}$$

Bu eşitlikte SS(Res)_p ele alınan p-değişkenli alt küme modelinden elde edilen kalıntı kareler toplamıdır ve s^2 ise bağımsız bilgi ya da daha genel olarak tüm bağımsız değişkenleri içeren modelden elde edilen σ^2 tahminidir. Model doğru olduğunda, kalıntı kareler toplamı $(n - p')\sigma^2$ nın yansız tahminidir. Bu durumda C_p burada yaklaşık olarak p' ye eşittir. Önemli bağımsız değişkenler modelden çıkartıldığında kalıntı kareler toplamı $(n - p')\sigma^2$ artı ihmal edilen değişkenlerin katkısını yansıtan pozitif miktardır. Bu durumda C_p nin p' den daha büyüktür.



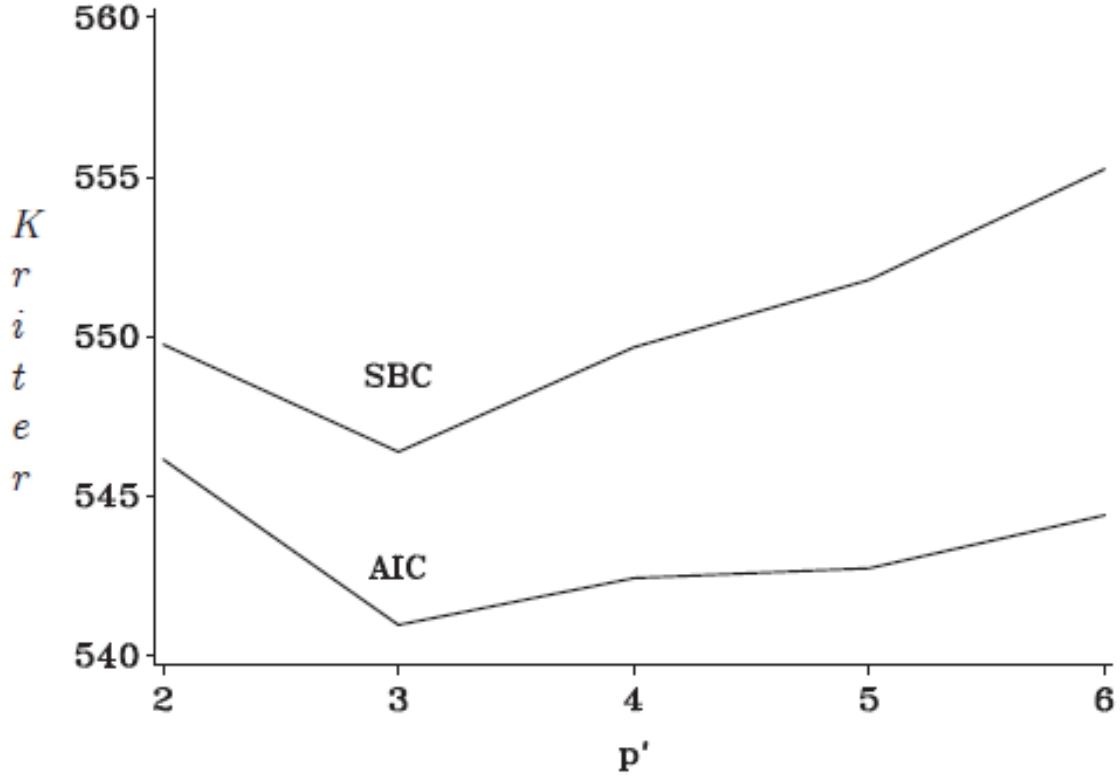
ŞEKİL 7.2. *Linthurst veri setinin C_p grafiği. Noktalı çizgi her bir alt küme büyüklüğü için $C_{p\,min}$i birleştirir. İki düz çizgi ise Hocking'in kriterine göre alt küme seçimi için referans çizgileridir.*

BİLGİ KRİTERİ (AIC) ŞU ŞEKİLDE HESAPLANMAKTADIR

Akaike (1969) Bilgi Kriteri (AIC) şu şekilde hesaplanmaktadır

$$\mathrm{AIC}(p') = n \ln(\mathrm{SS}(\mathrm{Res})_p) + 2p' - n \ln(n). \tag{7.4}$$

(Dikkat edilirse dökümandaki tüm logaritmik fonksiyonlar e tabanını kullanmaktadır.) $\mathrm{SS}(\mathrm{Res})_p$ bağımsız değişkenlerin sayısı arttıkça azalmaktadır. AIC deki ilk terim p' ile azalır. Halbuki, AIC deki ikinci terim p' ile artar ve modeldeki parameter sayısını artırmaya karşı bir penaltı vazifesi görmektedir. Böylece, bu tahmini elde etmek için kullanılan parameter sayısı ile tahmin kesinliği arasında bir ödünleşim verir. p' ye karşı çizilen $\mathrm{AIC}(p')$ grafiği, genelde minimum değeri gösterir ve uygun alt küme büyüklüğü için değer p' değeri ile belirlenir. $\mathrm{AIC}(p')$ burada minimum değerine erişecektir.



ŞEKİL 7.3. *Linthurst veri seti analizi için minimum AIC ve SBC değerleri p' ye karşı her bir alt küme büyüklüğü için çizilmiştir.*

ALT KÜME BÜYÜKLÜĞÜ SEÇİMİ İÇİN ANLAMLILIK DÜZEYLERİ

- Adımsal değişken seçim yöntemlerindeki F-giriş ve F-dur ya da denk anlamlılık düzeyleri tüm alt küme büyüklükleri gözetilmeden önce, seçim sürecini durdurmak için alt küme seçim kriteri olarak kullanılmaktadır.

- Bendel ve Afifi (1977) geleceğe dönük seçim için bir takım durdurma kurallarını karşılaştırmaktadırlar ve sabit anlamlılık düzeyine dayalı ardışık F- testini tercih etmiştir.

- Optimum giriş için anlamlılık düzeyi SLE = 0,15 ile 0,25 arasında değişmiştir.S

MODEL GEÇERLİLİĞİ

- Uyumlu regresyon (fitted regression) eşitliğinin geçerliliği, modelin anlamlı ve etkin olup olmadığının göstergesi ya da onayıdır.

- Uyumlu regresyonun veri seti ile hesaplanan eşitliğe denk olup olmamasının gösterimine denk değildir.

- Model geçerliliği, bağımsız bir veri setine karşı uyum regresyon eşitliğinin etkinliğinin değerlendirmesini gerektirir ve modeldeki güven bekleniyorsa önemlidir.

TABLO 7.4. *Gözlenen akış oranı, tahmin edilen akış oranı, su akım modelinin geçerlemesi için tahmin hatası. Sonuçlar akış oranı sırasına göre artan biçimde listelenmiştir* ($ft^3\ sec^{-1}$).

	Tahmin edilen P	*Gözlenen* Y	*Kestirim hatası* $\delta = P - Y$
	2,320	2,380	−60
	3,300	3,190	110
	3,290	3,270	20
	3,460	3,530	−70
	3,770	3,980	−210
	4,210	4,390	−180
	5,470	5,400	70
	5,510	5,770	−260
	6,120	6,890	−770
	6,780	8,320	−1,540
Ortalama	4,423	4,712	−289